HOW TO DESIGN DATA PROCESSING INPUT RECORDS FOR OPTIMUM RESULTS

John P. Kennedy

Card Design

Inefficiencies in machine processing resulting from poor card design can be measured in milliseconds or microseconds per record. Even when dealing with large files, this will usually add up to no more than a few minutes per run. If the run is repeated frequently, however, a few minutes or a few dollars difference per run may be significant. For a large library processing its circulation file daily, inefficiency resulting from poor card design and requiring a few extra milliseconds for processing each record could cost the library hundreds of dollars over the course of a year.

Usually a more serious consequence of mistakes in card design than increased processing time is increased time in coding and punching the data. Small differences in card layout can result in differences of several seconds per record in coding and punching. While clerical time is less expensive than computer time, the cost of a few extra seconds in coding the source document or punching the record will usually be more expensive than a few milliseconds of computer time. The most serious consequence of poor card design may be an increase in the number of errors made in preparing the input. Any change in source document or card layout which will result in fewer errors in the input data will probably prove to be economical even though it may increase processing time. The acceptable tolerance level for errors varies from one application to another, but in many library operations, errors eventually result in problems that require a considerable amount of professional time to solve. In cases in which the most convenient layout for coding and key punching is not the best layout for processing efficiency, priority should almost be given to the concenience of the persons producing the records rather than to the machine. The limitations and capabilities of the clerks recording and punching the data are more important considerations in card design than the

John P. Kennedy is Research Associate jointly with the Statistical Service Unit and the University Library, University of Illinois.

limitations and capabilities of the computer which will process data.

The most important consideration for the keypuncher is that she not have to skip from place to place on the source document in order to pick out items in the order in which they are to be punched. Source document and card layout should be planned so that items to be punched stand out for easy location and occur in sequence from top to bottom and left to right. It should be realized that any coding which must be done during keypunching will slow down the punching and will probably decrease accuracy. Right justification of fields should be avoided in punching. Right justification of numeric fields often makes processing more efficient, but it decreases punching speed and is one of the most common sources of errors in punching. In using unit record equipment right justification is often essential, but in computer input it is more efficient to let the computer take care of justification than to require the punch operator to do it. Another technique for facilitating the punching of long numeric fields and reducing the frequency of punching errors is to break the fields into shorter elements. It is difficult to keep long unbroken numbers in mind, and transposition errors are common in recording them. The practice of dividing long numbers into shorter elements is familiar in the telephone number and the Social Security number. It is especially advantageous if the elements can be meaningful. For example, in an accession number the first digits may represent the year of accession, or in an order number, part of the number may represent the fund on which an item is ordered.

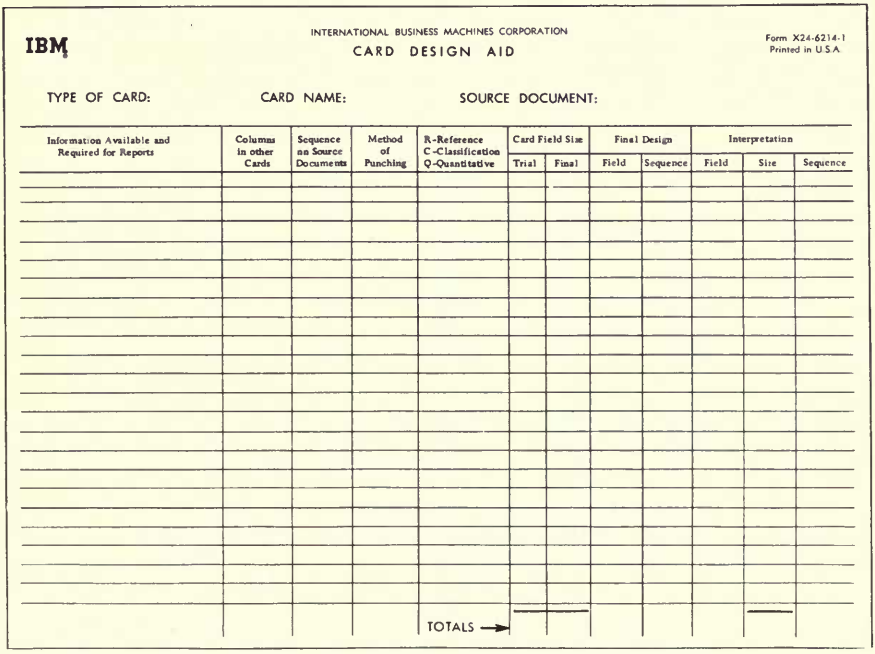
Works on forms design are available which detail other factors which can improve the efficiency and accuracy of recording information on forms of various types.[1] These may be especially helpful in designing dual purpose cards which are used for the original recording of data which will be keypunched into it. The additional factors which must be considered in designing cards for computer input (because of the nature of the machines) are not difficult to comprehend and require little technical knowledge of the computer. The one essential requirement is that the computer be able to distinguish different types of information. This is accomplished most often by the positions or fields into which different items are punched in the card and by codes which may be used to identify items and card types. Good card design requires the determination of the size and sequence of the fields for essential items of information so that keypunching and processing can be accomplished most efficiently.

There are four questions that must be answered in planning the card layout. These are:

1. What items of information must be punched?
2. How should each item be punched?
3. How large a field will be required for each item?
4. What sequence of fields will be most convenient?

The factors that will be relevant in answering each of these four questions follow.

What items of information must be punched? It is easiest to answer this question by beginning with the final products or outputs of the system. These will almost always be printed documents. In order to design the input records, it must first be determined what the desired output is; then proceed to list the items of information required for these outputs. In addition to items that actually appear on the final printed reports, these items must be analyzed to determine what additional data that does not appear may be needed for production. At this point it is advisable to include any information which may be useful even though it is not being recorded under existing routines. If it is likely that information will be needed in future operations, it is usually more economical to include it with the original input. Examples of information items which have not traditionally been recorded by libraries but are being considered for computer input at the Library of Congress are the language and color of books. The University of Toronto Library is including the thickness of books in its records for the Ontario New Universities even though thickness does not appear in the printed catalogs which are the main current output of the system.



IBM

INTERNATIONAL BUSINESS MACHINES CORPORATION
CARD DESIGN AID

Form X24-6214-1
Printed in U.S.A.

TYPE OF CARD: CARD NAME: SOURCE DOCUMENT:

Information Available and Required for Reports	Columns in other Cards	Sequence on Source Documents	Method of Punching	R-Reference C-Classification Q-Quantitative	Card Field Size		Final Design		Interpretation		
					Trial	Final	Field	Sequence	Field	Size	Sequence
				TOTALS →							

Figure 1
IBM Card Design Aid

Equipment and forms suppliers can provide various aids for card design. Figure 1 is a Card Design Aid supplied by IBM. It is convenient to use in recording the required data for card design. The first column of the form may be used to record the information items which are thought to be potentially useful.

Now one should eliminate as many of the listed items as possible. Often items which are ordinarily recorded are found to be redundant or useless. For example, some libraries have found it unnecessary to record both call number and author and title for circulation records. In a manual system the author and title provide a useful check. If the call number on a charge is illegible or incorrect, the author and title can be used to identify the book. With a mechanized system, this may not be needed.

After having eliminated any items that really are not needed, the remaining items should be examined if it is necessary for them to be punched. If processing is done regularly each day it is probably unnecessary to punch the date; the computer can add the date automatically to each record processed. The decimal point in Dewey classification numbers is another element that can be supplied automatically. There is no need to punch the point and carry it as an extra position in every record when the computer can easily supply it in printed output. Some information may be available through table look-ups. If particular departments or locations control specific funds, then it is unnecessary to punch both. The fund can be punched and a table used to determine which department or location the item goes to. Some information may be available in other files. If there is a vendor name and address file, then it will probably be necessary to identify a vendor only by a code in punching order and financial records. If there is a borrower name and address file; then only the borrower's identification number need be punched in circulation records. Finally, some information items may be calculated by the computer. If a library's loan periods are determined by borrower status and type of material, then it may be unnecessary to punch a due date for charged items since this can be calculated by the computer.

In the initial proposal for the conversion of one library's shelf list to magnetic tape, a three card set for each copy was called for. The first card would have been an author card, the second a title card, and the third would have given imprint, location, and order information. After review, the use of master cards for author and title with detail cards for each copy of the title was decided upon. The cards layouts are shown in Figure 2. Since this library has large numbers of copies of many of its titles, this simple change from three card sets to the use of master and detail cards resulted in important savings in punching and processing time. Examination of the data items included on the detail cards suggests that some

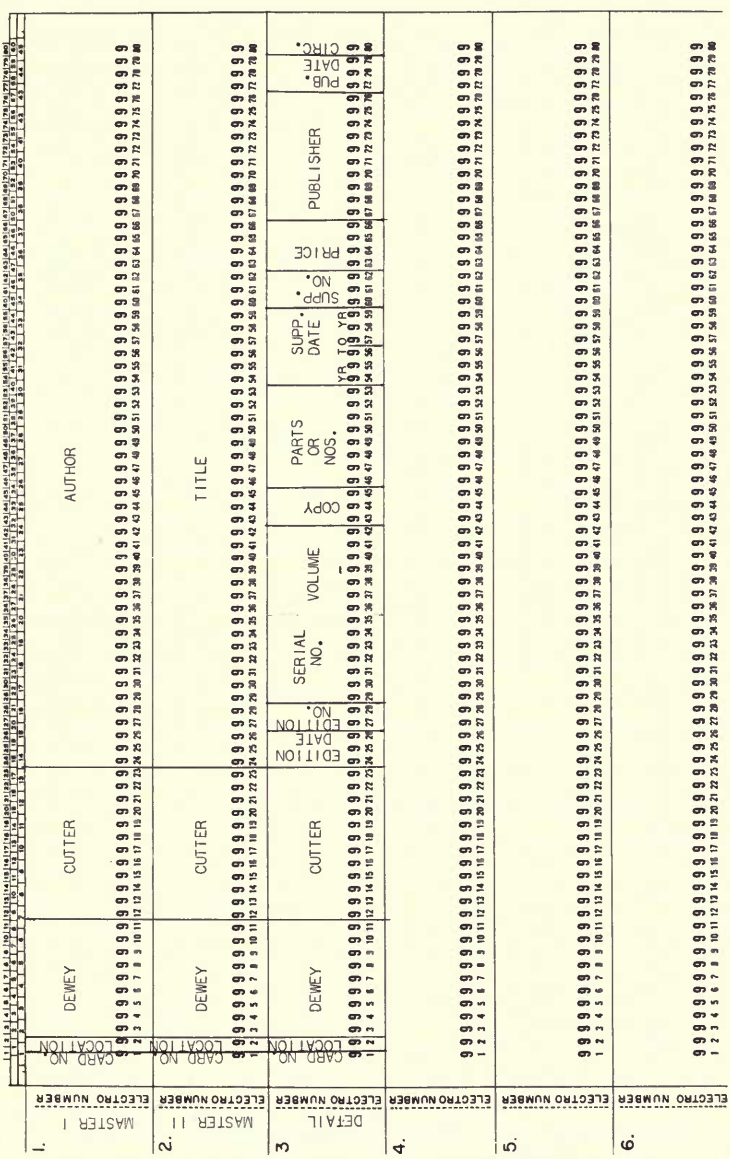


Figure 2
Multiple Card Layouts for Shelf List Data

unnecessary punching may still have been done. Several of the items on the detail card will be constant for all copies purchased at one time. It would be possible to punch the constant items, the inclusive copy numbers, and the inclusive serial or accession numbers for all copies purchased at one time. The necessity of including both copy number and serial number and both edition data and publication date might also be questioned.

The items which remain after such examination provide the answer to the first question, what items of information must be punched? The next question to be considered is how these items should be punched. Every alternative to keypunching should be considered for each item.

One possibility is that some items may be prepunched. Cards may be purchased with transaction numbers, order numbers, or accession numbers prepunched. In an acquisitions system, sets of cards may be punched by the computer when a book is initially ordered. Then by merely adding the appropriate status code, the cards can be used to update the processing file to reflect the current status of the book as reports are received or actions taken in processing it. In several serials systems now in operation, input cards for reporting the arrival of expected serial pieces are completely prepunched by the computer.

A second possibility is the automatic punching of items by gang punching, reproducing, or duplicating on a keypunch. If none of these methods for automatically punching the data are appropriate, then it may be wise to consider possible alternatives to the use of punched cards for input. Several libraries have decided on the use of optical scanning equipment for conversion of data from shelf list records. Both the University of Maryland and Southern Illinois University used optical mark readers in preparing book cards from their shelf lists. Johns Hopkins University, in converting more data from their shelf list, found it economical to have a service agency retype the records in a font which is readable by an optical scanner. Another alternative to the use of punched cards which should be considered is the use of punched paper tape. Whether or not this is a practical alternative will often be determined by the availability of equipment.

If it is decided that punched cards will be the best form for input, a final alternative to keypunching for some items may be mark sensing. Mark sensing may be advantageous in circumstances where a few short items of information are to be recorded at various stations in the library. The University of Missouri Library has found mark sensing useful in recording data for catalog statistics and for transfer and withdrawal statistics. In this type of use, the errors which are likely to occur in using mark sensing are not critical. At the Suffolk Cooperative Library System, mark sensing is used for adding price, discount, and classification number to previously

punched order cards. Errors would be critical in this application, but the director reports that in their circumstances mark sensing has been extremely accurate.[2]

The main alternatives to keypunching have been considered, and a decision should be reached as to the best method for punching each required item. The punching method selected for each item can be recorded on the Card Design Aid. The third question to be answered is how large a field is required for each item?

At this point it is advisable to determine whether other cards already in use include some of the same items of information. If so, the size of the field required is already determined. The card columns used for an item in other cards and the size of the field can be recorded on the Card Design Aid. It is also advisable to consider whether there are standardized layouts in use in the library or in other libraries or organizations which might be appropriate. A layout suggested by H. P. Luhn which is useful in many library applications is shown in Figure 3.[3] This is the format required for use in IBM's Keyword-In-Context (KWIC) and Selective Dissemination of Information (SDI) systems. It has been adopted with modifications for use in some procedures at the Pennsylvania State University Library, the University of California at Santa Cruz Library, the University of Illinois Library, and in the Urban Documentation and Retrieval Project. If it is likely that a library will wish to use either the KWIC or SDI system, then it should probably adopt a modification of this layout for those records which will be used in the KWIC or SDI system.

In addition to the advantages of standardization, the 60 column field for variable information has another merit. It makes it convenient to use two 60 character columns for printed output. This output format is efficient in that it utilizes most of the 132 print positions available on most printers. It also makes it possible to print the two columns, with a 4 print position separation on 8-1/2" $\times$ 11" paper, by reducing to about 58 or 60 per cent of original size. A number of libraries are photographically reducing printed output to this size for economy in reproduction. Reduction to only 58 or 60 per cent seems to be significantly easier to use and more pleasing to users than reduction to 50 per cent.

If the size of fields is not determined by the use of a standardized format or by the size allowed on other cards, then it will be necessary to determine the appropriate size. The field should be large enough to record the maximum number of characters that may be required for recording the item, unless the item can be shortened without loss of essential information. For some items such as date, order number, or borrower number, it is easy to establish the exact number of characters required. For other items, such as number of copies or price, practical maximums can easily be set. For such

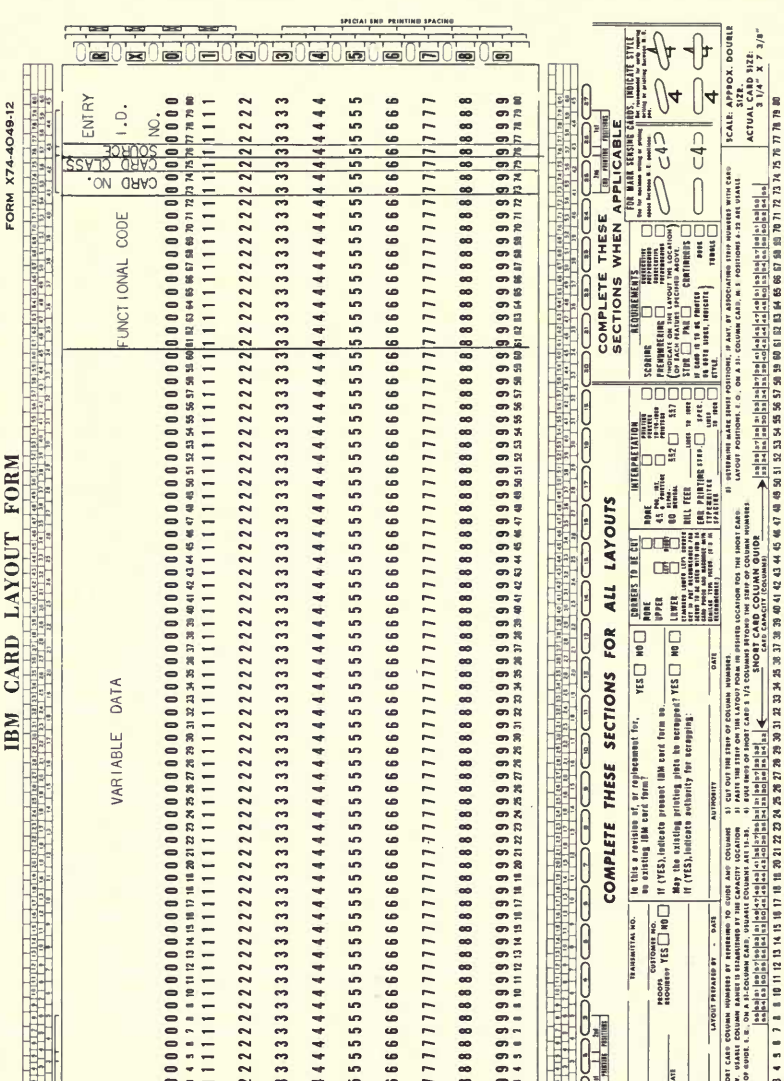


Figure 3
General Purpose Card Layout for Libraries
Developed by Luhn

items as entry and title, there are no logical maximums, and it is difficult to determine practical maximums. The best approach is to analyze a sample of the items. After such an analysis, one can decide at what point it will be acceptable to truncate the longer items or what size will accommodate a large enough percentage of the records that the remainder may be handled as exceptions.

Needs and practices vary so much from library to library it is impossible to suggest appropriate sizes for field such as call number, author and title. Among the card layouts of which the author has copies, there is a range in the number of columns allowed for call numbers from 16 to 51, with a fairly even distribution between the extremes.

The use of overly long fields in order to be certain to accommodate any item which may be encountered will increase the computer time for each pass. In effect, most of the computer time may be spent in processing blanks. In addition, available core capacity may be used inefficiently and tape handling time increased. In preparing a list of current periodical holdings at the University of Illinois, it was found that 90 per cent of the entries were less than 80 characters and 98 per cent were less than 120 characters. A title field of 256 characters would have been required, however, to accommodate the longest entry. The use of fixed length fields large enough for the maximum size of every item will often increase the size of records by several times and will increase computer time correspondingly.

If it has been decided that truncation of items is unacceptable and that it will not be efficient to deal with oversize items as exceptions, it may be advantageous to use variable length records. Input of variable length items is usually handled by using fixed fields on the cards but allowing for a variable number of cards. Cards then usually contain only one item or part of an item apart from reference data to identify the card. The type of item on each card is indicated by card numbers and codes for card type. It is also possible to identify variable length items by the use of flags or codes and by their sequence.

The use of variable length records decreases the size of the records and the file and therefore decreases the input and output time in each pass of that file. On the other hand, it means increased steps in processing and therefore makes programming more difficult and usually increases processing time. H. N. Laden and T. R. Gildersleeve, in a recent book on system design, give a good discussion of the factors that should determine the choice of fixed length or variable length records.[4] Maximum efficiency is achieved by a good balance of input-output time and processing time. If the run is input-output limited, the use of variable length records will likely improve overall time since it will decrease input-output time and increase

processing time. If the run is computer limited, the use of fixed length records will be preferable. The exact way of balancing depends on machine characteristics such as buffering capabilities and timing, but the following conditions suggest the use of variable length records:

1. The amount of variability in record size is extensive.
2. Among the variable items, the frequency of short items is relatively great.
3. The file contains a large volume of records and will be processed frequently.
4. The activity of the file is low.

You will notice that these characteristics are frequently found with the files used in library procedures in which full bibliographic descriptions are required. In such procedures, the use of variable length records will often be advantageous. If the advantages do not seem clear, however, it is probably best to use fixed length records. The difficulties in programming and problems in sorting for files of variable length records may be more serious than anticipated. Even if variable length records are utilized, it will probably be necessary to establish maximum sizes for the items.

After the size of each item has been determined and recorded, the total number of columns required can be summed. In some applications it is highly desirable that only one card be used for each unit record. For example, in a serials checking system there would be many problems and opportunities for error if an issue could be represented by more than one arrival card. In such applications, if the sum of the field sizes exceeds 80, it will probably be necessary to decrease the size of some fields. The simplest and most common method is further to truncate fields such as the field for an abbreviated title. In some cases it may be possible to shorten several fields by providing a common overflow field.

Occasionally it is possible to shorten fields without loss of information by more coding of data. An item such as a date may be represented in as few as four positions, as opposed to the thirteen or more which we usually use in writing it, yet still appear in conventional form in printed outputs. In most cases where an item is limited to a small number of possibilities, coding may be utilized. If the number of possibilities is less than 10, single numeric characters may be used. If it exceeds 10 but not 12, as for months of the year, numeric and zone punches may be utilized. Use of alphabetic and special characters permits a greater number of possibilities. Through the use of multi-punching, hundreds or even thousands of possibilities can be coded in two columns. In some cases, it is possible to overpunch control codes in columns used for numeric data or to combine in a single column the coding for two variables with

only a few possible conditions. Such heavy coding should be avoided in most circumstances, however, since coding of the data is slower, the probability of error in coding and punching becomes greater, and computer editing for accuracy becomes difficult or impossible. Some additional techniques that may be used for reducing the number of columns required in special circumstances are suggested in the IBM manual, Form and Card Design.[5]

After determining the size of the field required for each item, the final question to be answered is what order of items will be most convenient? If the format of the source document is fixed, the sequence of items on the source document is the most important factor in determining the sequence for the card. If the source document has not been designed or is to be redesigned, several factors may be considered in determining the best sequence for items on the source document and card. If other cards used by the library include some of the same items of information, the same card columns should usually be used for these items. This may facilitate programming by making it possible to copy parts of file descriptions or subroutines from existing programs. Keypunching may be facilitated since the operator becomes familiar with a single location in which an item is punched, and fewer different program cards may be required. Another factor that may be considered is the desired output format. It is better to arrange items so that a minimum of reformatting is required in order to produce the desired output. In many library applications, one card is used for the input of one complete line of information in the same format that it will be listed in computer printed catalogs or processing lists. This usually insures a minimum of programming effort and computer time.

The nature of the information may also influence the order of items. The Card Design Aid shown in Figure 1 provides a column for the classification of items as Reference, Classification, or Quantitative. Reference items, such as order number or transaction number, identify the other data on the card. Classification items such as call number, borrower number, account number, and status codes, are used to group records for reports. Quantitative data includes items such as quantities and prices. Conventionally, cards are arranged with reference data to the left and quantitative data to the right. This classification is not very useful in most library applications, but the practice of placing fixed length reference type items to the left is helpful. To most of us, it seems natural to place the fields for items such as call number, author and title to the left, and reference items such as identification numbers, card types, and card numbers to the right. If, however, reference items which are usually of fixed length and must always be punched are placed at the left of the card and items which may vary in length or may be left blank are placed to the right, punching is expedited. After punching

the required data, the operator can touch the reject button and then find her place on the next document while the card is being ejected and the next card registered. Location of fields to be automatically duplicated at the left or right end of the card also adds to the time available to the operator for finding her place on the next document.

Finally, the sequence of items should be planned with consideration for machine limitations and capabilities. In the use of equipment which reads input serially such as the IBM 357 units which are used in several library circulation systems, any blank columns should be at the right of the card so that machine time is not wasted in reading blank columns. If several items are to be used in one sorting operation, it is desirable that these items be in adjacent fields with the major element to the left and the minor element to the right. In some cases, the order of items may determine whether it will be possible to chain instructions and thereby save a little processing time.

The location of items on dual use cards presents several additional factors for consideration. In planning the layout for dual cards, it is important to consider the visibility of items to be punched. Information on the card may be partially concealed by the punch housing unit or by the pressure arm of the keypunch. It may also be necessary to position essential written information so that it will not be obliterated by the punches. The use of dual cards has advantages in many applications. The card may include information such as signatures which cannot be punched or which it is unnecessary to punch. For example, in a circulation procedure, a tabulating card form might be completed by the borrower. It would be necessary to punch only a few items into the card for machine processing, but the complete record including signature and address would be available for overdue procedures. The use of dual cards often eliminates the need for typing the card to a source document through reference items and the necessity of maintaining two files.

With the determination of the sequences of items on the card, all of the information needed for the final layout is at hand. A number of card layout forms are available for use in recording the layout and for ordering custom printed cards. IBM supplies a multiple card layout form for sets of cards, a general purpose card layout form, a dual card layout form, and a number of other layout forms for more unusual types of cards. Figure 2 shows a multiple card layout form, and Figure 3 shows a general purpose card layout form. These forms include a number of guides and scales useful in drafting the layout. Instructions for the use of these forms are included in the manual, Form and Card Design.

REFERENCES

1. Marien, Ray. Marien on Forms Control. Englewood Cliffs, N.J., Prentice-Hall, 1962. Sadauskas, Wallace B. Manual of Business Forms. New York, Office Publications, 1961.

2. Curley, Walter W. "The Data Processing Program in Operation at the Suffolk Cooperative Library System, Patchogue, New York," see this volume.

3. Luhn, Hans Peter. General Rules for Creating Machinable Records for Libraries and Special Reference Files. (Form No. 225-1487). Yorktown Heights, N.Y., IBM Corp., Advanced Systems Development Division, 1960.

4. Laden, H. N., and Gildersleeve, T. R. System Design for Computer Applications. New York, Wiley, 1963, pp. 92-97.

5. International Business Machines Corporation. Data Processing Techniques: Form and Card Design. (Form C20-8078). White Plains, N.Y., IBM Corp. Technical Publications Department, 1961, p. 8.